Networks and Grids for Science and Global Virtual Organizations

Harvey B. Newman

NASA ESSAAC Meeting
UCSD (and Rio de Janeiro)
February 18 2004

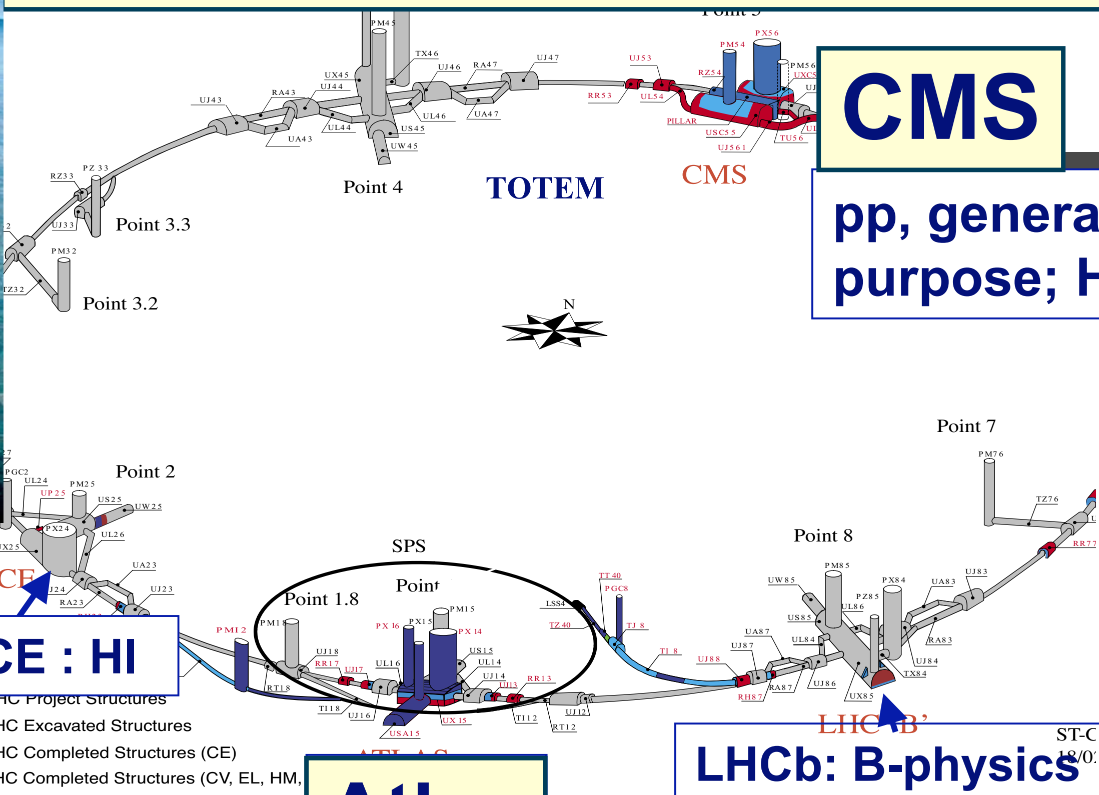
The Challenges of Next Generation Science in the Information Age

Petabytes of complex data explored and analyzed by 1000s of globally dispersed scientists, in hundreds of teams

- Flagship Applications
  - High Energy & Nuclear Physics, AstroPhysics Sky Surveys: TByte to PByte “block” transfers at 1-10+ Gbps
  - eVLBI: Many real time data streams at 1-10 Gbps
  - BioInformatics, Clinical Imaging: GByte images on demand
- HEP Data Example:
  - From Petabytes in 2003, ~100 Petabytes by 2007-8, to ~1 Exabyte by ~2013-5.
- Provide results with rapid turnaround, coordinating large but limited computing and data handling resources, over networks of varying capability in different world regions
- Advanced integrated applications, such as Data Grids, rely on seamless operation of our LANs and WANs
  - With reliable, quantifiable high performance

Large Hadron Collider (LHC) CERN, Geneva: 2007 Start

- pp $\sqrt{s}$ =14 TeV L=10^{34} cm^{-2} s^{-1}
- 27 km Tunnel in Switzerland & France



CMS

pp, general purpose; HI

ALICE : HI

Atlas

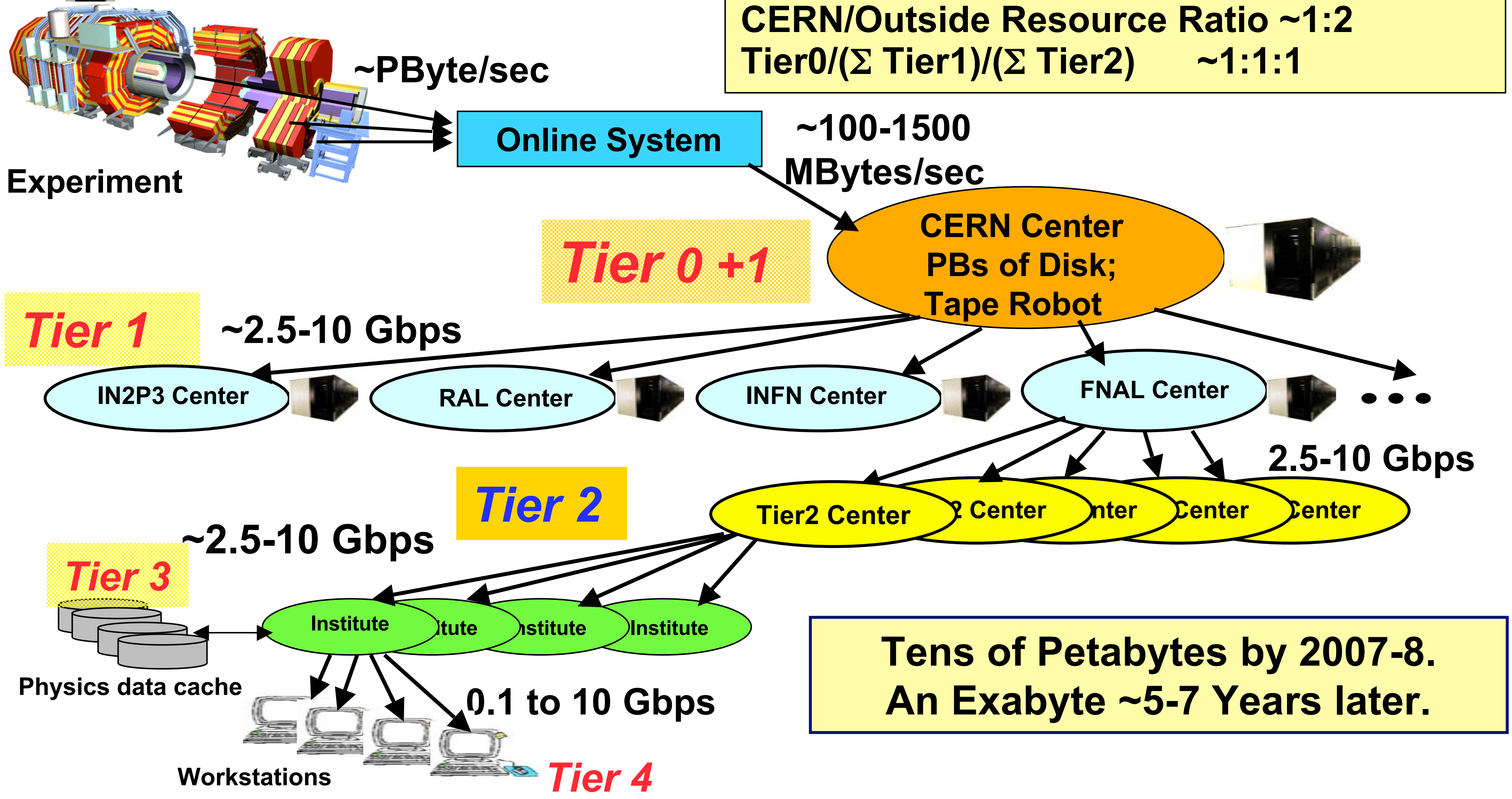
LHCb: B-physics

First Beams: April 2007

Physics Runs: from Summer 2007

LHC Data Grid Hierarchy: Developed at Caltech

CERN/Outside Resource Ratio ~1:2
Tier0/(Σ Tier1)/(Σ Tier2) ~1:1:1

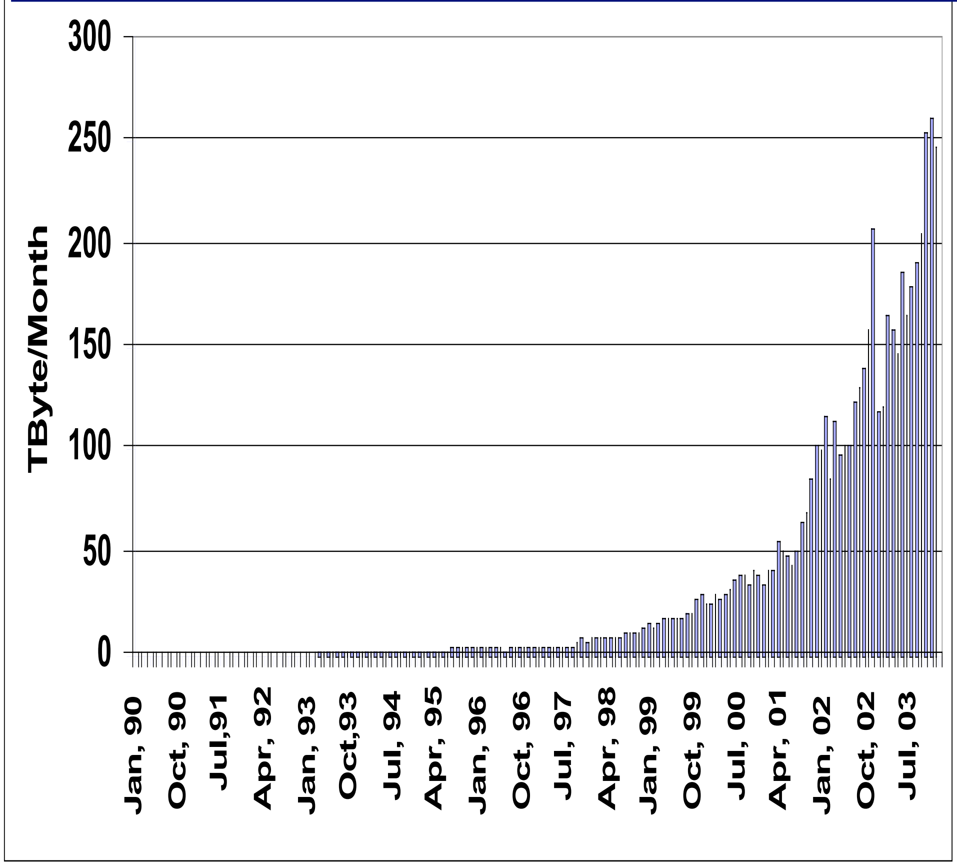


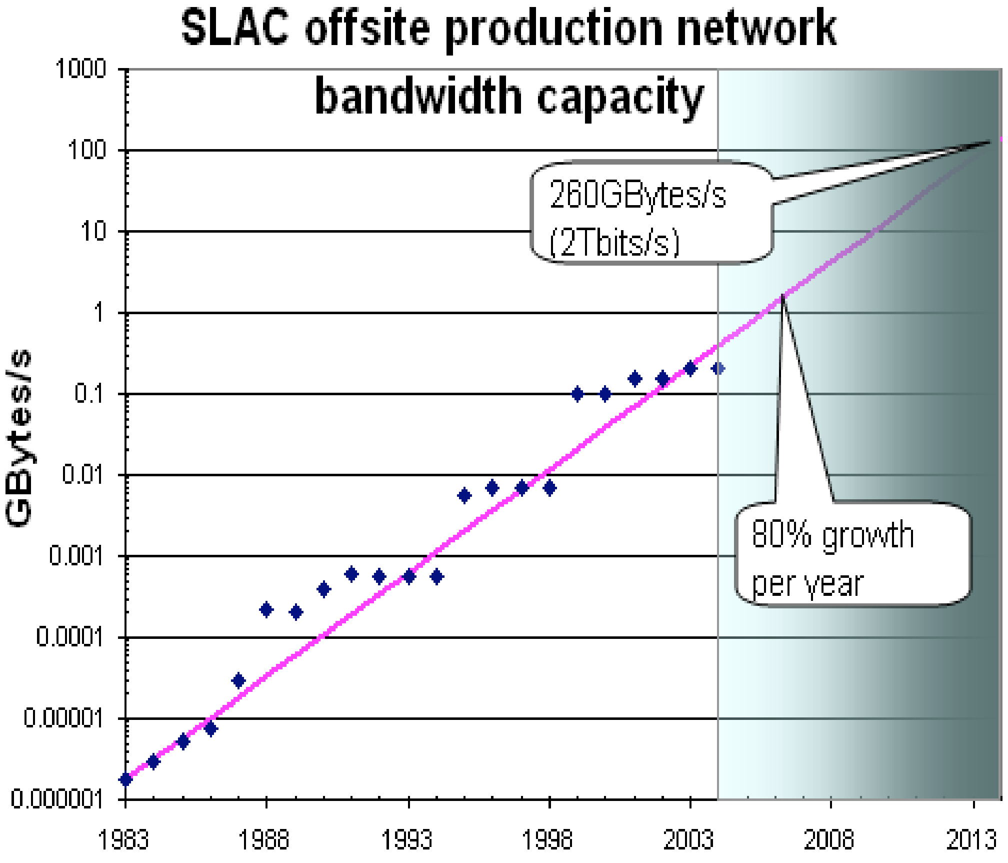
Tens of Petabytes by 2007-8.
An Exabyte ~5-7 Years later.

Emerging Vision: A Richly Structured, Global Dynamic System

History of Bandwidth Usage – One Large Network; One Large Research Site

ESnet Accepted Traffic 1/90 – 1/04
Exponential Growth Since ’92;
Annual Rate Increased from 1.7 to 2.0X
Per Year In the Last 5 Years





SLAC Traffic ~300 Mbps; ESnet Limit
Growth in Steps: ~ 10X/4 Years
Projected: ~2 Terabits/s by ~2014

Fall 2003: Transatlantic Ultraspeed TCP Tranfers
Throughput Achieved: X50 in 2 years

Terabyte Transfers by the Caltech-CERN Team:

- ◆Nov 18: 4.00 Gbps IPv6 Geneva-Phoenix (11.5 kkm)
- ◆ Oct 15: 5.64 Gbps IPv4 Palexpo-L.A. (10.9 kkm)
  - ❒ Across Abilene (Internet2) Chicago-LA, Sharing with normal network traffic
  - ❒ Peaceful Coexistence with a Joint Internet2-Telecom World VRVS Videoconference

Nov 19: 23+ Gbps TCP: Caltech, SLAC, CERN, LANL, UvA, Manchester

Juniper,
HP
Level(3)
Telehouse

HENP Major Links: Bandwidth Roadmap (Scenario) in Gbps

Year	Production	Experimental	Remarks
2001	0.155	0.622-2.5	SONET/SDH
2002	0.622	2.5	SONET/SDH DWDM; GigE Integ.
2003	2.5	10	DWDM; 1 + 10 GigE Integration
2005	10	2-4 X 10	? Switch; ? Provisioning
2007	2-4 X 10	~10 X 10; 40 Gbps	1st Gen. ? Grids
2009	~10 X 10 or 1-2 X 40	~5 X 40 or ~20-50 X 10	40 Gbps ? Switching
2011	~5 X 40 or ~20 X 10	~25 X 40 or ~100 X 10	2nd Gen ? Grids Terabit Networks
2013	~Terabit	~MultiTbps	~Fill One Fiber

Continuing the Trend: ~1000 Times Bandwidth Growth Per Decade; We are Rapidly Learning to Use Multi-Gbps Networks Dynamically

HENP Lambda Grids: Fibers for Physics

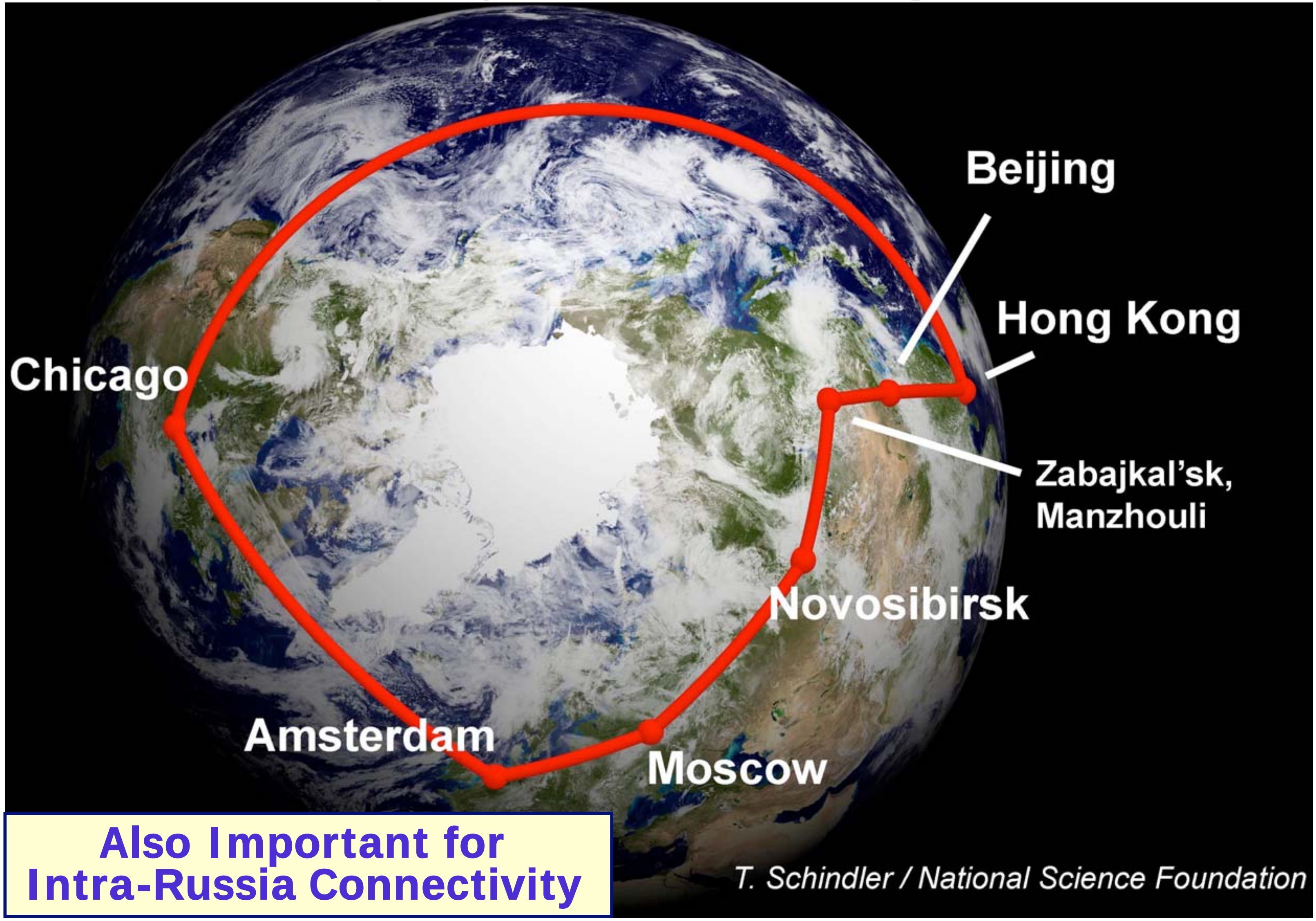
- Problem: Extract "Small" Data Subsets of 1 to 100 Terabytes from 1 to 1000 Petabyte Data Stores
- Survivability of the HENP Global Grid System, with hundreds of such transactions per day (circa 2007) requires that each transaction be completed in a relatively short time.
- Example: Take 800 secs to complete the transaction. Then

Transaction Size (TB)	Net Throughput (Gbps)
1	10
10	100
100	1000 (Capacity of Fiber Today)

- Summary: Providing Switching of 10 Gbps wavelengths within ~2-4 years; and Terabit Switching within 5-8 years would enable "Petascale Grids with Terabyte transactions", to fully realize the discovery potential of major HENP programs, as well as other data-intensive research.

GLORIAD: Global Optical Ring (US-Ru-Cn)

“Little Gloriad” (OC3) Launched January 12; to OC192 in 2005



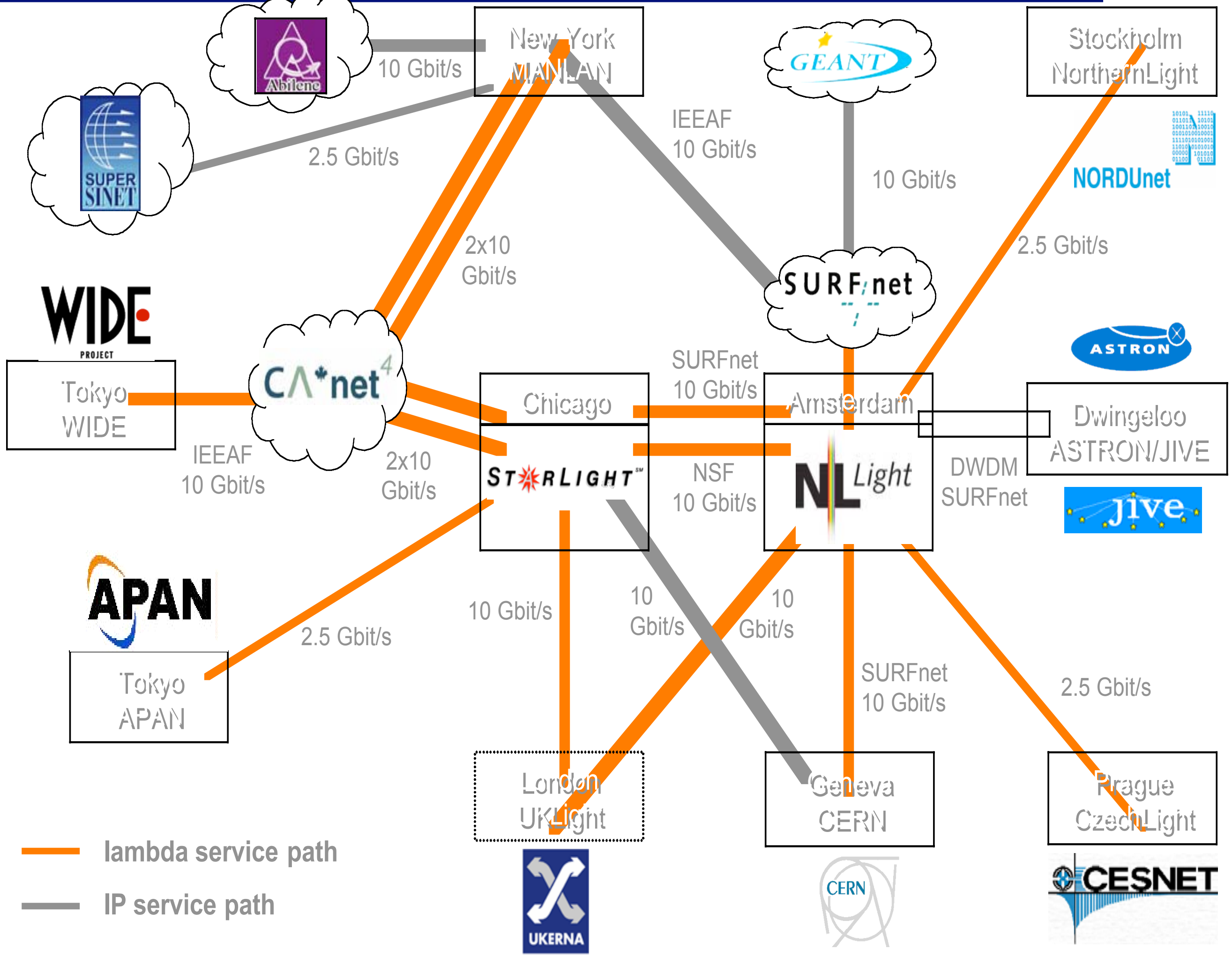
T. Schindler / National Science Foundation

Also Important for Intra-Russia Connectivity

GLIF: Global Lambda Integrated Facility

"**GLIF** is a World Scale Lambda based Lab for Application and Middleware development, where Grid applications ride on dynamically configured networks based on optical wavelengths ...

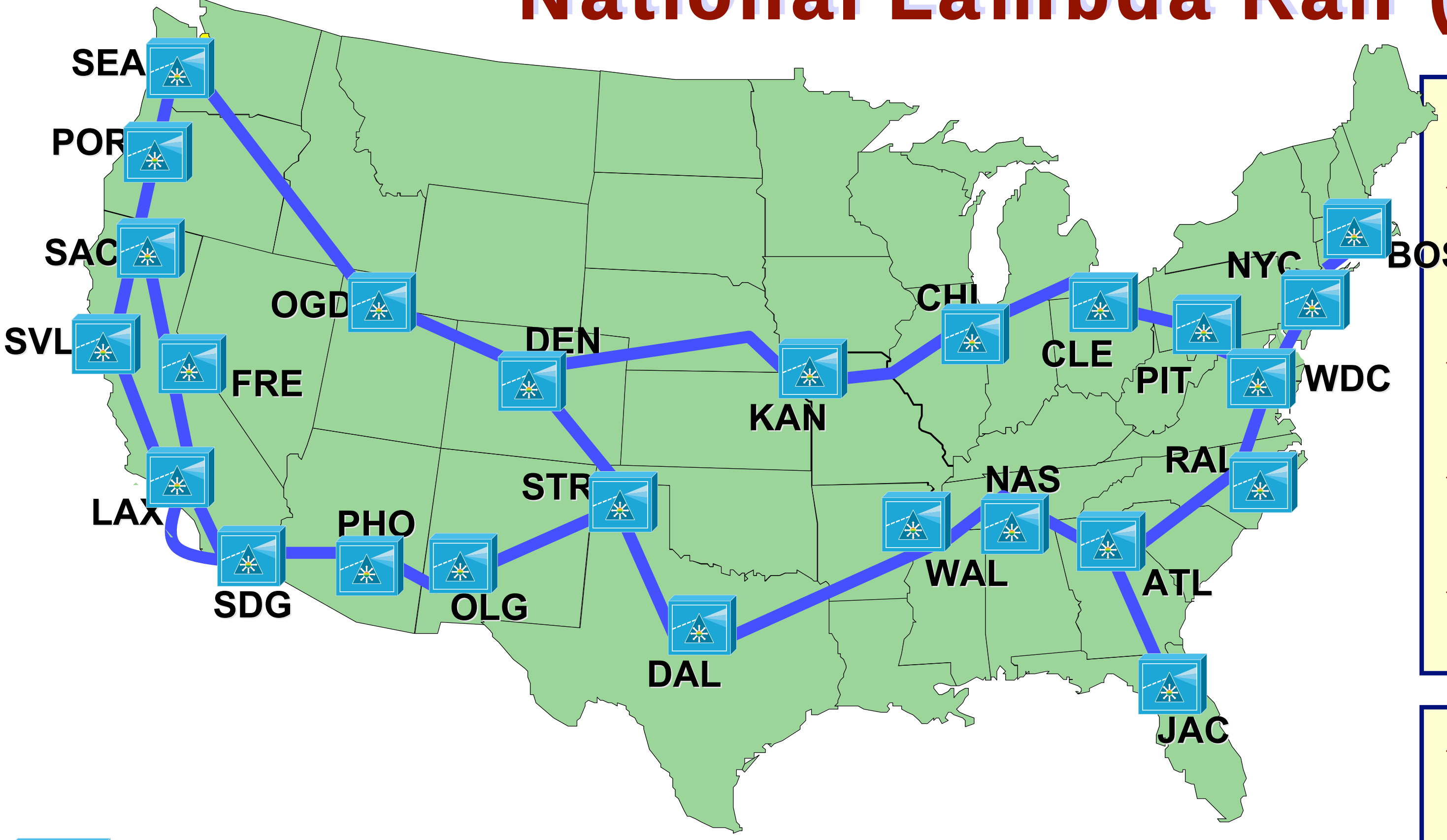
GLIF will use the Lambda network to support data transport for the most demanding e-Science applications, concurrent with the normal best effort Internet for commodity traffic."



10 Gbps Wavelengths For R&E Network Development Are Prolifering, Across Continents and Oceans

Transition beginning now to optical, multi-wavelength Community owned or leased fiber networks for R&E

National Lambda Rail (NLR)



NLR

- ◆Coming Up Now
- ◆Initially 4 10G Wavelengths
- ◆Full Footprint Ops by 3Q or 4Q04
- ◆Internet2 HOPI Initiative (w/HEP)
- ◆To 40 10G Waves in Future

- ◆ Regional Dark Fiber Initiatives in 18 U.S. States

▣ 15808 Terminal, Regen or OADM site

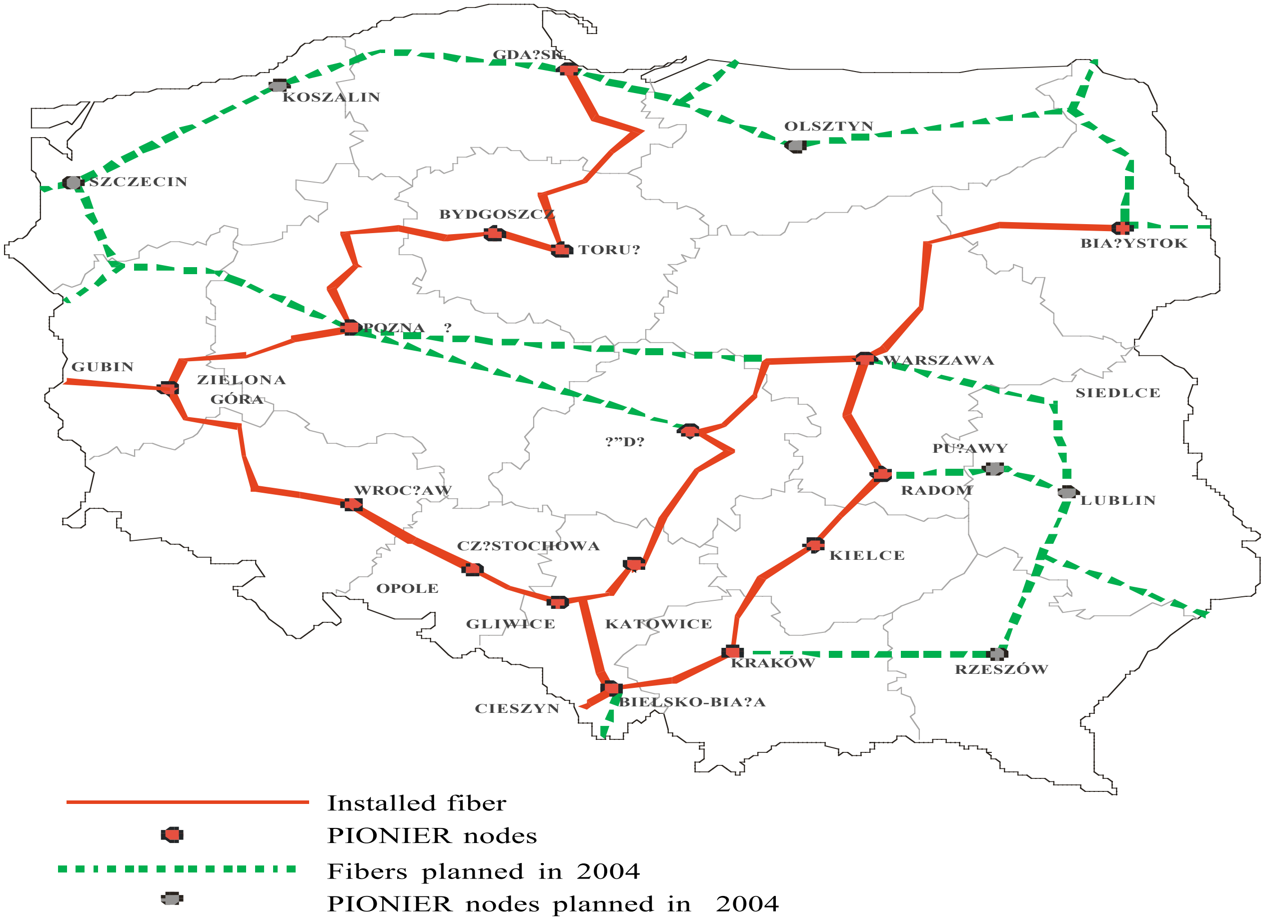
▬ Fiber route

Dark Fiber in Eastern Europe
Poland: *PIONIER* Network

2650 km Fiber Connecting 16 MANs; 5200 km and 21 MANs by 2005

Support

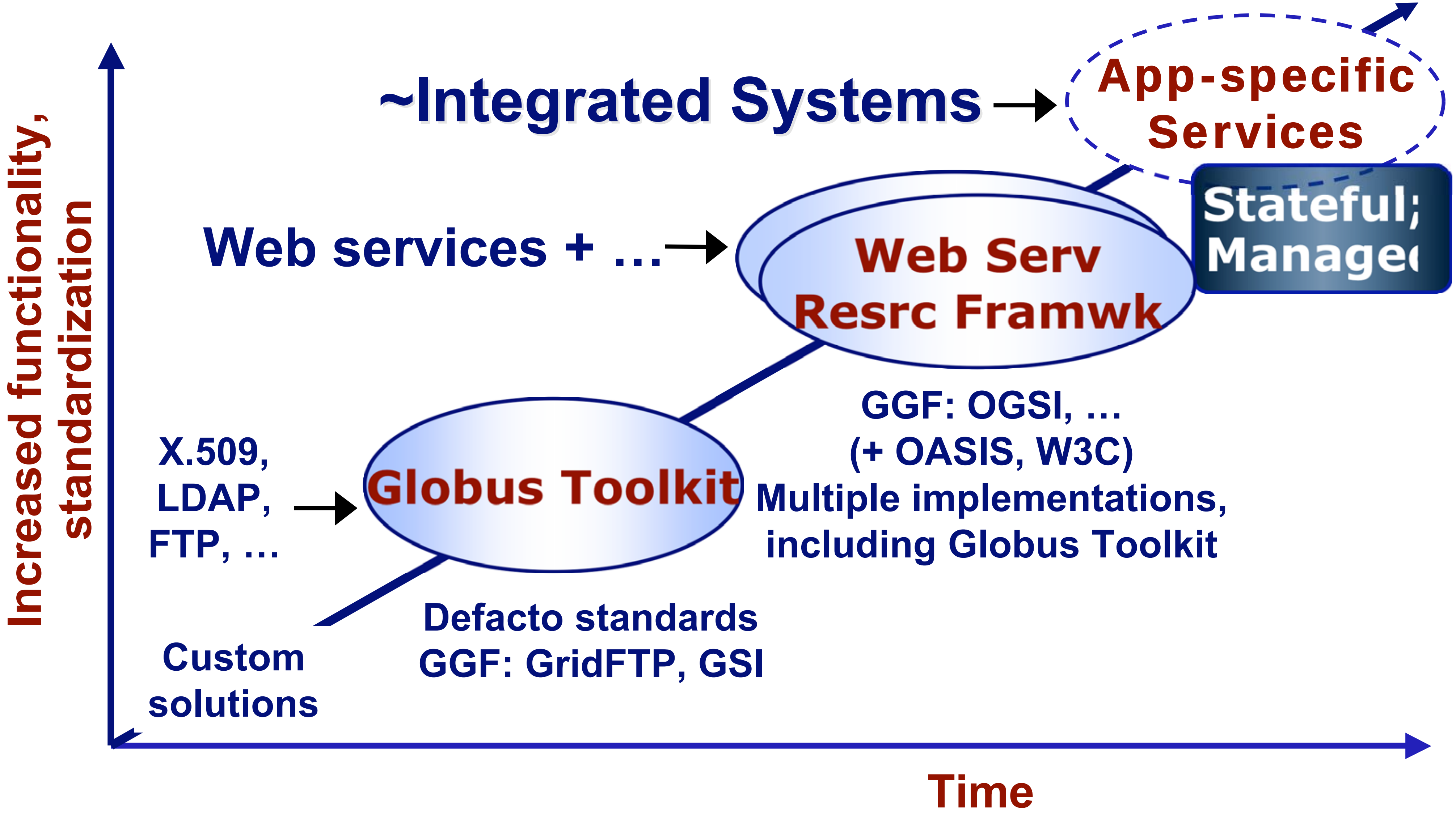
- Computational Grids Domain-Specific Grids
- Digital Libraries
- Interactive TV
- Add'l Fibers for e-Regional Initiatives



Classical, HENP Data Grids, and Now Service-Oriented Grids

- The original Computational and Data Grid concepts are largely stateless, open systems: known to be scalable
  - Analogous to the Web
- The classical Grid architecture has a number of implicit assumptions
  - The ability to locate and schedule suitable resources, within a tolerably short time (i.e. resource richness)
  - Short transactions with relatively simple failure modes
- HENP Grids are *Data Intensive & Resource-Constrained*
  - 1000s of users competing for resources at 100s of sites
  - Resource usage governed by local and global policies
  - Long transactions; some long queues
- HENP → Stateful, End-to-end Monitored and Tracked Paradigm
  - Adopted in OGSA, Now WS Resource Framework

The Move to OGSA and then Managed Integration Systems



Increased functionality, standardization

Time

Custom solutions

X.509, LDAP, FTP, …

Globus Toolkit

Defacto standards GGF: GridFTP, GSI

Web services + …

Web Serv Resrc Framwk

GGF: OGSI, … (+ OASIS, W3C) Multiple implementations, including Globus Toolkit

~Integrated Systems

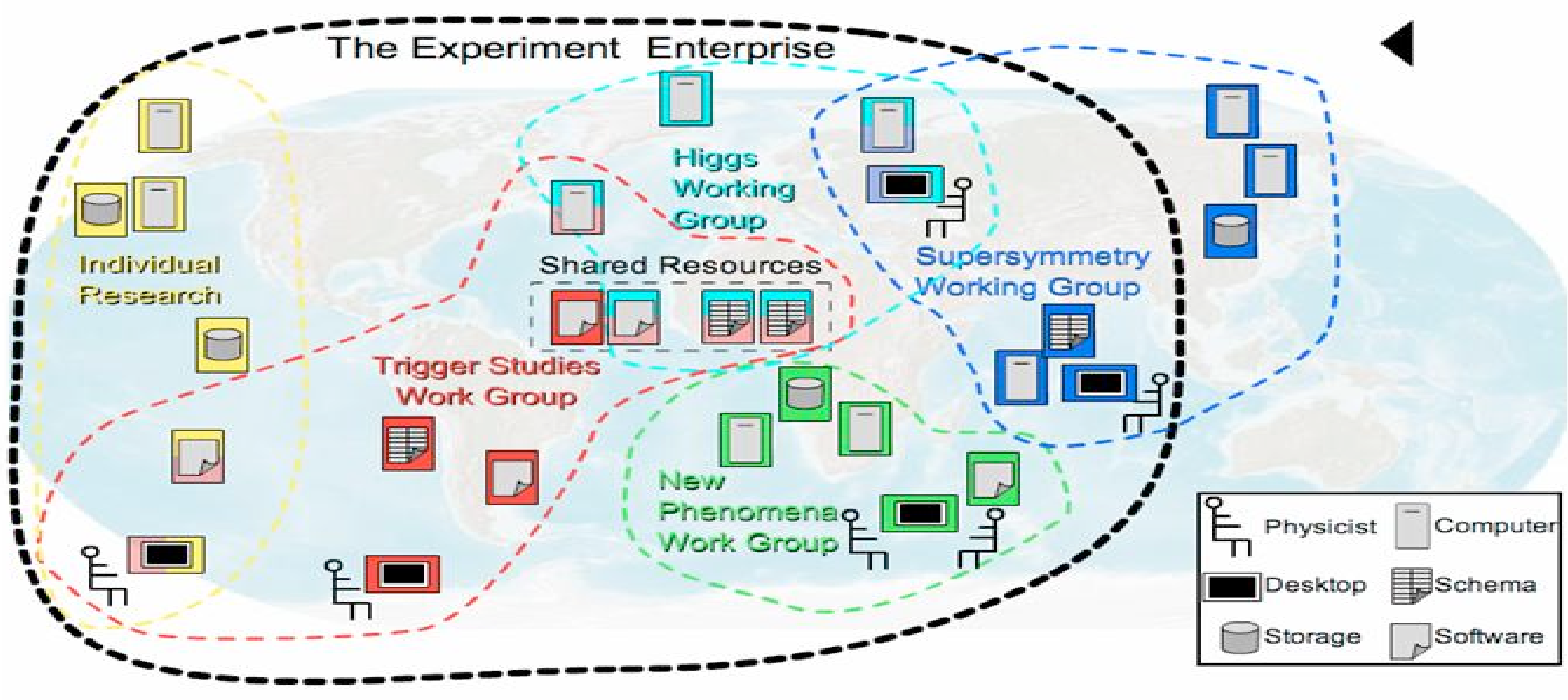
App-specific Services

Stateful; Manage

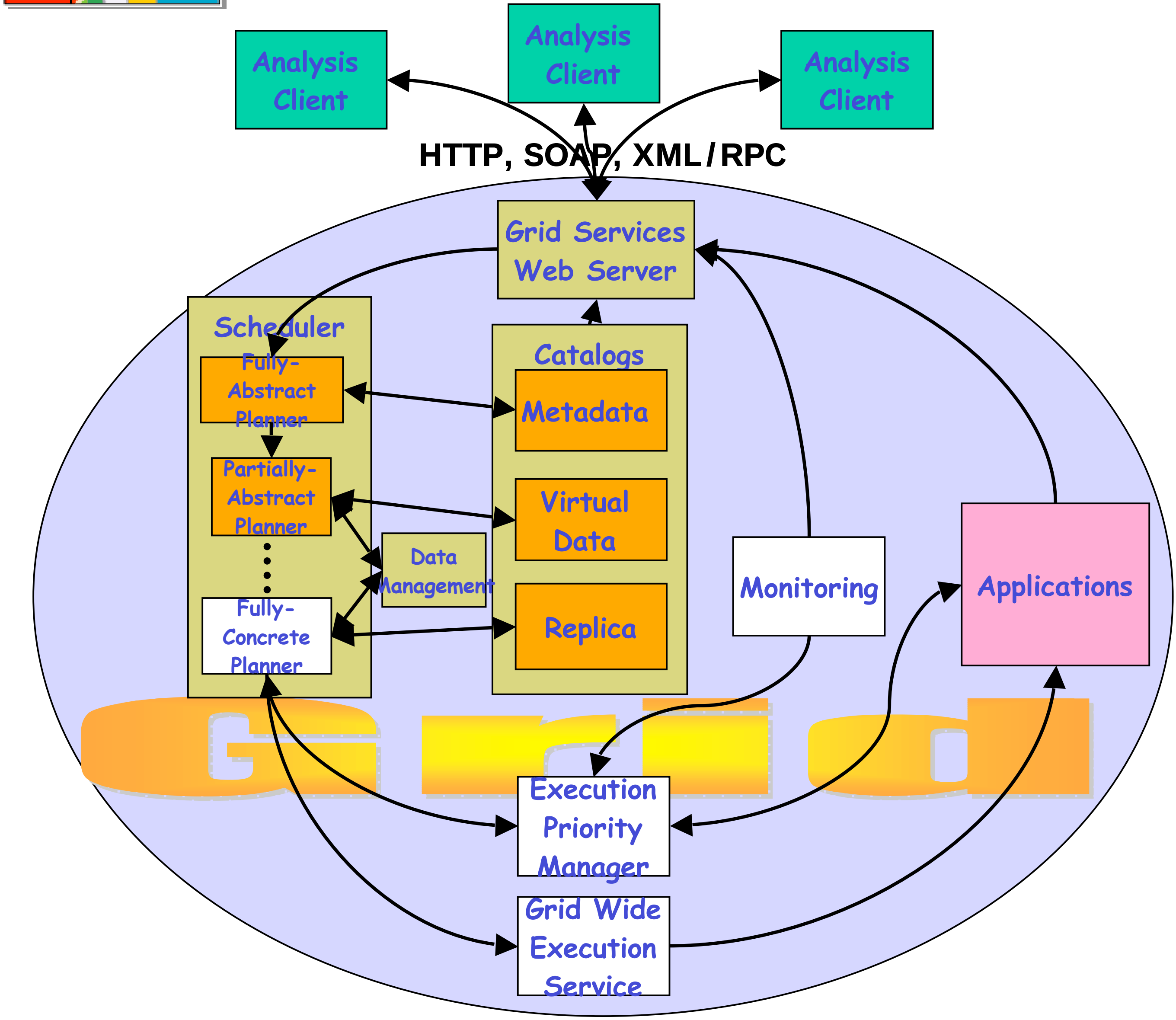
The Grid Analysis Environment (GAE)

The GAE: key to “success” or “failure” for physics & Grids in the LHC era:

- 100s - 1000s of tasks, with a wide range of computing, data and network resource requirements, and priorities



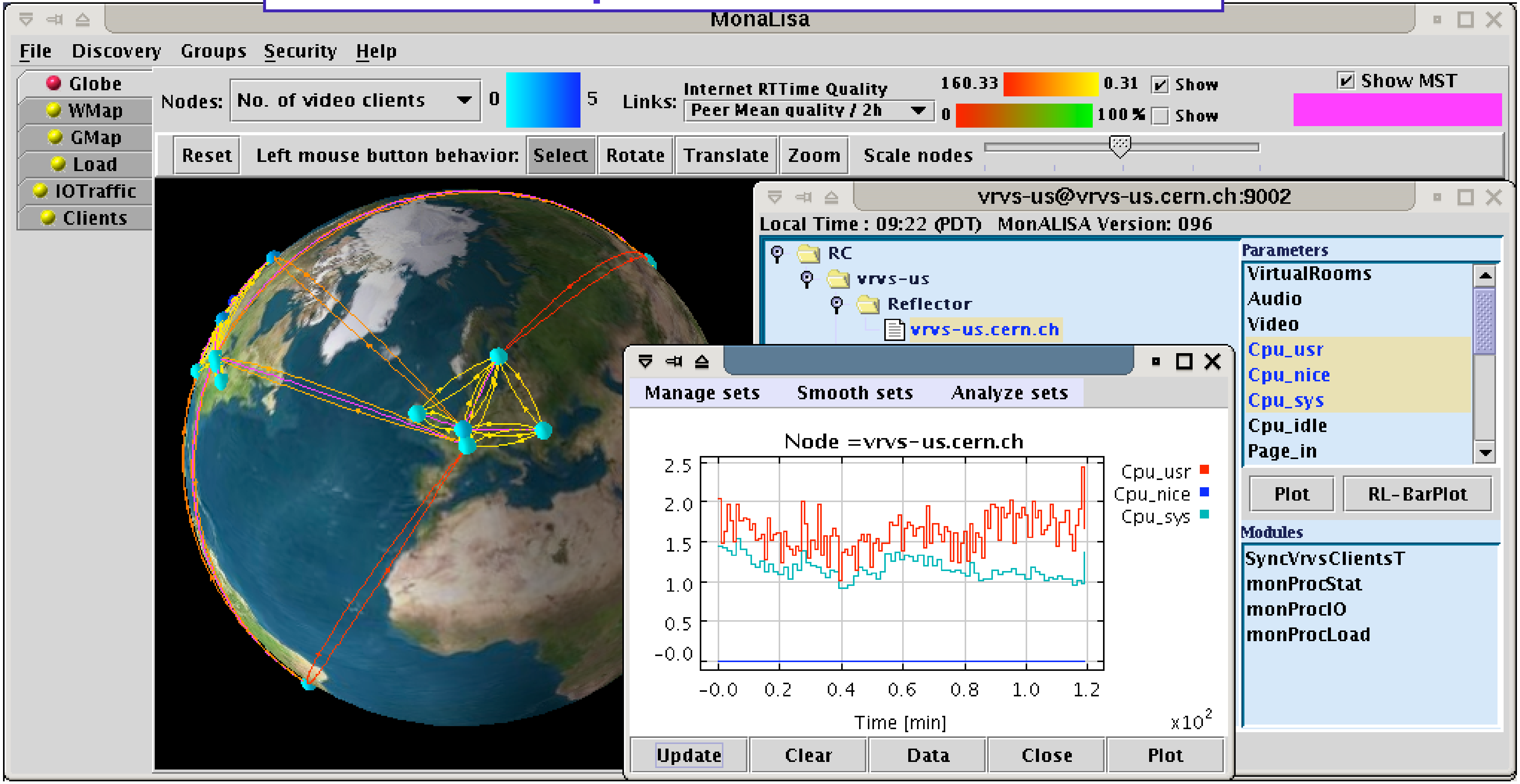
GAE Architecture



- Analysis Clients talk standard protocols to the “Grid Services Web Server”, a.k.a. the Clarens data/services portal.
- The Clarens portal hides the complexity of the Grid Services from the client, but can expose it in as much detail as req’d for e.g. monitoring.
- Key features: *Global* Scheduler, Catalogs, Monitoring, and Grid-wide Execution service. Clarens servers form a ***Global Peer network***.

Managing Global Systems: Dynamic Scalable Services Architecture

MonALISA: http://monalisa.cacr.caltech.edu



MonaLisa

Nodes: No. of video clients

Links: Internet RTTime Quality

Peer Mean quality / 2h

☑ Show MST

Left mouse button behavior: Select

vrvs-us@vrvs-us.cern.ch:9002

Local Time : 09:22 (PDT) MonALISA Version: 096

- RC
  - vrvs-us
    - Reflector
      - vrvs-us.cern.ch

Parameters

- VirtualRooms
- Audio
- Video
- Cpu_usr
- Cpu_nice
- Cpu_sys
- Cpu_idle
- Page_in

Modules

- SyncVrvsClientsT
- monProcStat
- monProcIO
- monProcLoad

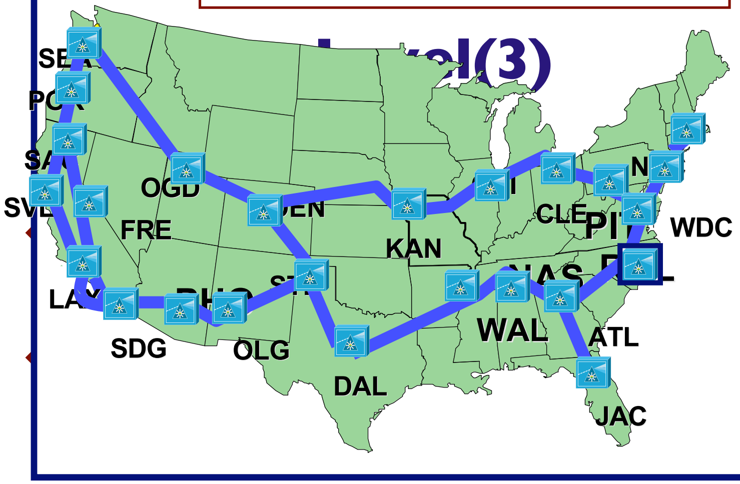
Node =vrvs-us.cern.ch

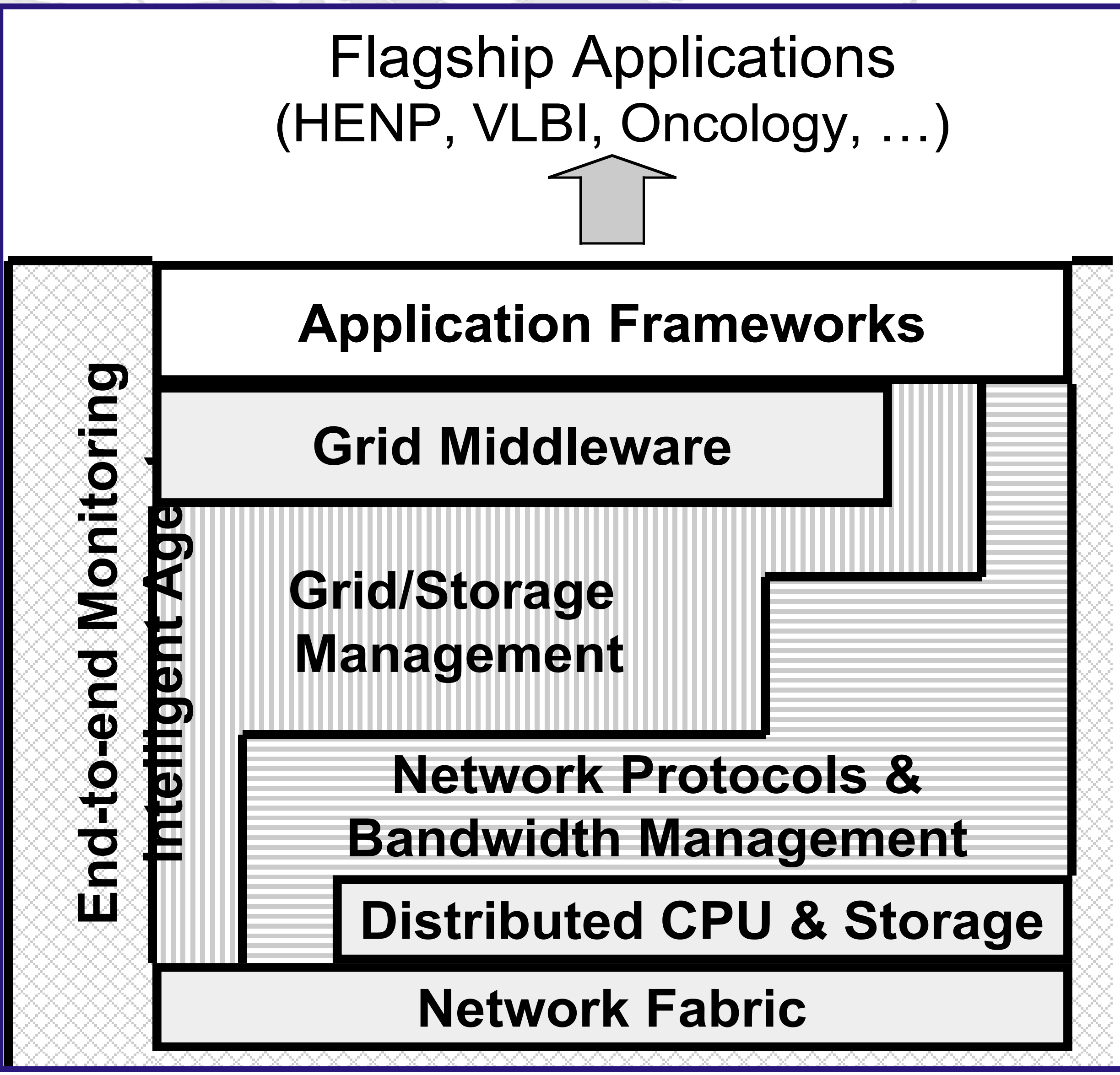
UltraLight Collaboration: http://ultralight.caltech.edu

- Caltech, UF, FIU, UMich, SLAC,FNAL, MIT/Haystack, CERN, UERJ(Rio), NLR, CENIC, UCAID, Translight, UKLight, Netherlight, UvA, UCLondon, KEK, Taiwan

Level(3)

National Lambda Rail





Flagship Applications
(HENP, VLBI, Oncology, …)

End-to-end Monitoring

Intelligent Age

Application Frameworks

Grid Middleware

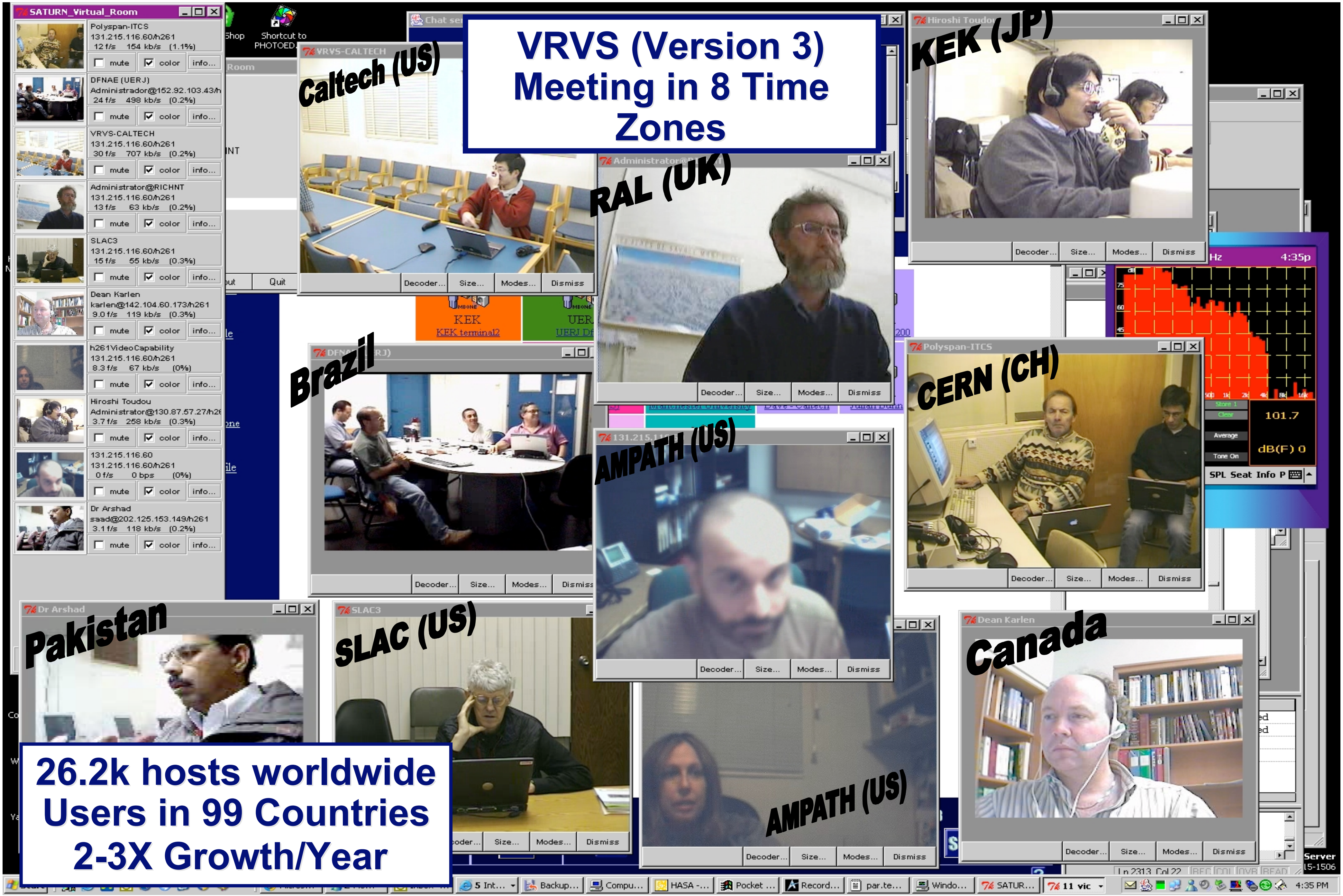
Grid/Storage Management

Network Protocols & Bandwidth Management

Distributed CPU & Storage

Network Fabric

VRVS (Version 3) Meeting in 8 Time Zones



```
Polyspan-ITCS
131.215.116.60/h261
 12 f/s    154 kb/s   (1.1%)
□ mute   ☑ color   info...
DFNAE (UERJ)
Administrador@152.92.103.43/h
 24 f/s    498 kb/s   (0.2%)
□ mute   ☑ color   info...
VRVS-CALTECH
131.215.116.60/h261
 30 f/s    707 kb/s   (0.2%)
□ mute   ☑ color   info...
Administrator@RICHNT
131.215.116.60/h261
 13 f/s     63 kb/s   (0.2%)
□ mute   ☑ color   info...
SLAC3
131.215.116.60/h261
 15 f/s     55 kb/s   (0.3%)
□ mute   ☑ color   info...
Dean Karlen
karlen@142.104.60.173/h261
 9.0 f/s   119 kb/s   (0.3%)
□ mute   ☑ color   info...
h261VideoCapability
131.215.116.60/h261
 8.3 f/s    67 kb/s    (0%)
□ mute   ☑ color   info...
Hiroshi Toudou
Administrator@130.87.57.27/h26
 3.7 f/s   258 kb/s   (0.3%)
□ mute   ☑ color   info...
131.215.116.60
131.215.116.60/h261
 0 f/s      0 bps      (0%)
□ mute   ☑ color   info...
Dr Arshad
saad@202.125.153.149/h261
 3.1 f/s   118 kb/s   (0.2%)
□ mute   ☑ color   info...
```

26.2k hosts worldwide
Users in 99 Countries
2-3X Growth/Year

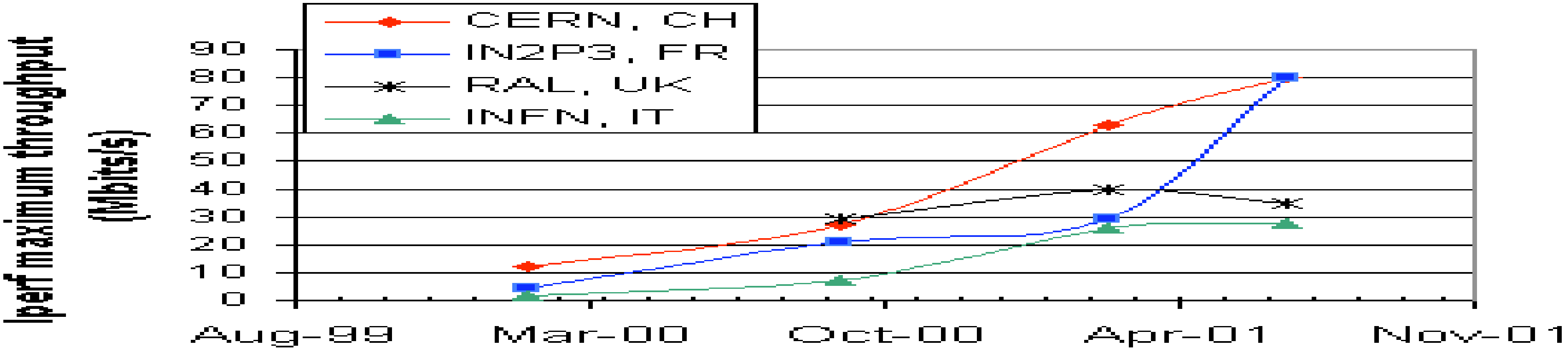
Networks, Grids and HENP

- Network backbones and major links used by HENP experiments are advancing rapidly
  - To the 2.5-10G range in < 2 years; much faster than Moore's Law
- HENP is learning to use long distance 10 Gbps networks effectively
  - 2003 Developments: to 5.6+ Gbps flows over 11,000 km
- Transition to a community-owned or leased fibers for R&E has begun in some areas [us, ca, nl, pl, cz, sk] or is considered [de, ro; IEEAF]
- End-to-end Capability is Needed, to Reach the Physics Groups:
  - Removing Regional, Last Mile, Local Bottlenecks and Compromises in Network Quality are now *On the critical path, in all world regions*
- *Digital Divide: Network improvements are especially needed in SE Europe, Latin America, China, Russia, Much of Asia, Africa*
- Work in Concert with Internet2, Terena, APAN, AMPATH; DataTAG, the Grid projects and the Global Grid Forum

Recommendation 1: Work on the Digital Divide from Several Perspectives

- Work on Policies and/or Pricing: pk, in, br, cn, SE Europe, …
  - ❒ Share Information: Comparative Performance and BW Pricing
  - ❒ Find Ways to work with vendors, NRENs, and/or Gov'ts
  - ❒ Exploit Model Cases: e.g. Poland, Slovakia, Czech Republic
- Inter-Regional Projects
  - ❒ South America: CHEPREO (US-Brazil); EU ALICE Project
  - ❒ GLORIAD, Russia-China-US Optical Ring
  - ❒ Virtual SILK Highway Project (DESY): FSU satellite links
- Help with Modernizing the Infrastructure
  - ❒ Design, Commissioning, Development
  - ❒ Provide Tools for Effective Use: Monitoring, Collaboration
- Participate in Standards Development; Open Tools
  - ❒ Advanced TCP stacks; Grid systems
- Workshops and Tutorials/Training Sessions
  - ❒ For Example: Rio DD and HEPGrid Workshop, February 2004
- Raise General Awareness of the Problem; Approaches to Solutions

HEP is Learning How to Use Gbps Networks Fully: Factor of ~500 Gain in Max. Sustained TCP Thruput in 4 Years, On Some US+Transoceanic Routes



- ◆ 9/01 105 Mbps 30 Streams: SLAC-IN2P3; 102 Mbps 1 Stream CIT-CERN
- ◆ 5/20/02 450-600 Mbps SLAC-Manchester on OC12 with ~100 Streams
- ◆ 6/1/02 290 Mbps Chicago-CERN One Stream on OC12
- ◆ 9/02 850, 1350, 1900 Mbps Chicago-CERN 1,2,3 GbE Streams, 2.5G Link
- ◆ 11/02 [LSR] 930 Mbps in 1 Stream California-CERN, and California-AMS FAST TCP 9.4 Gbps in 10 Flows California-Chicago
- ◆ 2/03 [LSR] 2.38 Gbps in 1 Stream California-Geneva (99% Link Utilization)
- ◆ 5/03 [LSR] 0.94 Gbps IPv6 in 1 Stream Chicago- Geneva
- ◆ TW & SC2003: 5.65 Gbps (IPv4), 4.0 Gbps (IPv6) in 1 Stream Over 11,000 km